Instance-level matching

Adam Sanchez, Tatiana Lesnikova, Jérôme David, Jérôme Euzenat

NSFC-61261130588

ANR-12-IS02-002-01

Lindicle

Linked data interlinking in a cross-lingual environment
跨语言环境中语义链接关键技术研究
Liage des données dans un environnement interlingue

D3.2 Instance-level matching

Coordinator: Adam Sanchez-Ayte
With contributions from: Tatiana Lesnikova, Jérôme David, Jérôme Euzenat

Quality reviewer:	Zhichun Wang
Reference:	Lindicle/D3.2/v2
Project:	Lindicle ANR-NSFC Joint project
Date:	September 13, 2016
Version:	2
State:	final
Destination:	public

Executive Summary

Ontology matching is a difficult task but it may becomes even more difficult when ontologies are defined in different natural languages because it becomes impossible to rely on class and relation names.

In order to overcome this problem, it is tempting to resort to instance-based ontology matching which takes advantage of common instances expressed with respect to the two ontologies. However, such instances may not be readily available.

This report first describes a general framework in which instance-based ontology matching may be expressed to take advantage of linked instances instead of common instances. It first reduces the ontologies to their backbone which is the strict necessary relations of this ontology which are taken into account to compute and update instance-based matching.

Finding subsumption and equivalence correspondences are extracted by measuring if their sets of instances tends to be included in each others. These sets of instances are expressed here with respect to links.

The comparison between sets of instances is parameterised by a specific measure. We specifically presents the entropic intensity measure which combines a statistical test and an inclusion index based on Shannon entropy.

This type of instance-based ontology matching is evaluated on real world data sets (DBPedia, XLore, GeoLinkeData and GeoSpecies). They mix English language and Chinese language sources. This evaluation showed promising results with a precision over 60% and practicable runtime.

DOCUMENT INFORMATION

Project number	ANR-NSFC Joint project	**Acronym**	Lindicle
Full Title	跨语言环境中语义链接关键技术研究 Linked data interlinking in a cross-lingual environment Liage des données dans un environnement interlingue		
Project URL	`http://lindicle.inrialpes.fr/`		
Document URL			

Deliverable	**Number**	3.2	**Title**	Instance-level matching
Work Package	**Number**	3	**Title**	Cross lingual ontology matching based on aligned cross lingual human-readable knowledge bases

Date of Delivery	**Contractual**	M42	**Actual**	2014-07-14
Status	final		final ☒	
Nature	prototype ☐ report ☒ dissemination ☐			
Dissemination level	public ☒ consortium ☐			

Authors (Partner)	Tatiana Lesnikova, Jérôme David, Jérôme Euzenat			
Resp. Author	**Name**	Adam Sanchez-Ayte	**E-mail**	Adam-Sanchez-Ayte@inria.fr
	Partner	INRIA		

Abstract (for dissemination)	This paper describes precisely an ontology matching technique based on the extensional definition of a class as a set of instances. It first provides a general characterisation of such techniques and, in particular the need to rely on links across data sets in order to compare instances. We then detail the implication intensity measure that has been chosen. The resulting algorithm is implemented and evaluated on XLore, DBPedia, LinkedGeoData and Geospecies.
Keywords	instance-based matching, ontology alignment

Version Log			
Issue Date	**Rev No.**	**Author**	**Change**
2015-04-20	1	J. Euzenat	Created template
2015-06-03	2	A. Sanchez	Added text about measures to compute degree of similarity and subsumption
2015-06-19	3	A. Sanchez	Extended version containing more details about the matching process (version 1)
2016-06-20	4	A. Sanchez	Introduced text from paper
2016-06-21	5	A. Sanchez	Added text for experimental evaluation
2016-06-21	6	A. Sanchez	Added appendix
2016-06-28	7	A. Sanchez	Updated tables about datasets and results
2016-06-30	8	A. Sanchez	Modified graphic about ontology backbone materialization
2016-07-07	9	A. Sanchez	Added table about materializations+various corrections
2016-07-08	10	A. Sanchez	Added comments about quality of Xlore data
2016-09-02	11	J. Euzenat	Added executive summary
2016-09-13	12	A. Sanchez	Taken quality control comments into account

Table of Contents

1. Introduction

Instance-based alignment extraction is a technique that exploits identity links between instances from different datasets. It is useful for practical applications where schema elements like classes have obscure, ambiguous names, different concept granularities or incomparable categorization. It is based on two key ideas: a) that the more significant the overlap of common instances of two classes is, the more related these classes are, b) the real semantics of a concept is often better defined by the actual instances assigned to the class than by annotations like the class name.

Since, nowadays, large volumes of data are produced and made available on the web [Margara et al. 2014], the computation of ontology alignments from their extension have received particular interest in the last years. Although all approaches assume similarity between classes based on the overlap of their set of instances, they differ in the use of measures to compute relatedness. Whereas some techniques follow classical approaches to compute pairwise Jaccard similarity [Correndo et al. 2012; Thor et al. 2007; Parundekar et al. 2010], new approaches use machine learning algorithms like Locality-Sensitive Hashing (LSH) [Duan et al. 2012; Zong et al. 2015] and Markov Random Field [Wang et al. 2008] to yield results with a very low error rate.

Instance-based alignment is the most suitable technique to find class-to-class correspondences between Xlore and DBpedia because the Xlore dataset contains classes and instances which URI identifiers use numeric values and it contains a relevant amount of *owl:sameAs* links pointing out to DBpedia.

The remainder of the report is organized as follows. In Section 2, we introduce the instance-based ontology matcher used to extract the alignment. To this end, the notion of ontology backbone and the implication intensity measure are presented. Section 3 presents our experimental evaluation. Results are discussed in Section 4. We conclude and outline possibilities for future work in Section 5.

2. Instance-based ontology matching

In order to present the alignment extraction process, we first introduce the notion of ontology backbone (§2.1). Then we present the generic instance-based ontology matching process (§2.2) that we instantiate with a specific implication intensity measure (§2.3).

2.1 Ontology backbone

Our starting point is to consider any data set together with its ontology as a set of triples or RDF graph G. Because we are focusing on instance-based matching we will concentrate on what we call the *ontology backbone*. In the following, classes are identified by the letters c, d, ... and instances by the letters i, j,... eventually subscripted.

From an RDF graph G the ontology backbone will be made of all subsumption (rdfs:subClassOf) and membership (rdf:type) statements entailed by this graph. So, this backbone can be further decomposed into two disjoint sets:

$$S(G) = \{(c,d) | G \models \langle c, \mathsf{rdfs{:}subClassOf}, d\rangle\}$$
$$M(G) = \{(i,c) | G \models \langle i, \mathsf{rdf{:}type}, c\rangle\}$$

Such a backbone may be obtained from G by materialising its classification. Below we will freely use $(c \sqsubset d) \in S$ to denote that $(c,d) \in S$ and $(d,c) \notin S$. By the same token, we will openly use $(p \sqsubset q) \in T$ to denote that $(p,q) \in T$ and $(q,p) \notin T$

We further denote the backbone ontology O corresponding to G, that we will call *ontology* from now on assuming G implicit, as the tuple

$$\langle C(G), I(G), S(G), M(G)\rangle$$

such as:

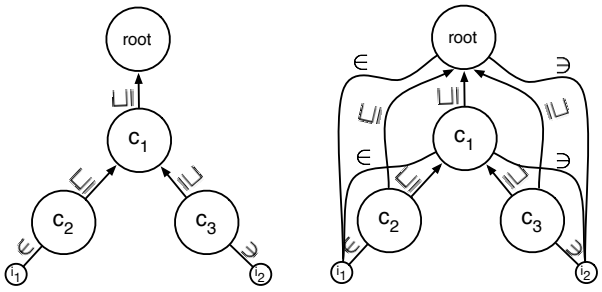
$$C(G) = \{c | G \models \langle c, \mathsf{rdf{:}type}, \mathsf{rdfs{:}Class}\rangle\}$$
$$I(G) = \{i | \exists c \in C(G); \langle i, c\rangle \in M(G)\}$$

A simplified and partial view of the evolving scenario to study is described by Figure 2.1(a). The intensional level of a ontology $O = \langle C, I, S, M\rangle$ is circumscribed to a taxonomy whereas the extensional level is represented by the membership relations between classes and their instances. The membership relation between an instance $i_1 \in I$ and a class $c_2 \in C$ is represented as a pair $\langle i_1, c_2\rangle \in M$. Directed arrows between classes denote subsumption relations in S. For instance, $\langle c_2, c_1\rangle \in S$.

Figure 2.1(b) shows that some membership relations may also be inferred based on the semantics of RDFS using the subsumption relationships between classes. For instance, $\langle i_2, c_1\rangle \in M$ but it is entailed from $\langle i_2, c_3\rangle \in M$ and $\langle c_3, c_1\rangle \in S$.

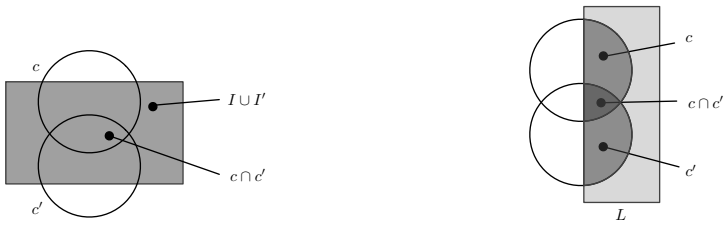
2.2 Instance-based matching

Let $O = \langle C, I, S, M\rangle$ and $O' = \langle C', I', S', M'\rangle$ be two ontologies. An alignment A between O and O' is a set of correspondences $\langle c, c', r, m\rangle$ where $c \in C$, $c' \in C'$, $r \in \{\sqsubseteq, \sqsupseteq, \equiv\}$ and $m \in [0\ 1]$ is a confidence degree associated to the correspondence. The goal of ontology matching is to find such alignments between two ontologies.



(a) The initial ontology O (b) The materialization of O

Figure 2.1: Two representations of the same ontology.



(a) (b)

Figure 2.2: Subsets to compute a measure m between c and c'.

Instance-based ontology matchers rely on ontology backbones in order to generate alignments. In order to do so, they use instances which are common to the two ontologies and compute measures between each pair of classes. Then, they extract an alignment from the measured values.

The measure used by such matcher usually relies on the following cardinalities (see Figure 2.2(a)):

1. $|c \cap c'|$: the number of instances belonging to both classes c and c'.
2. $|c|$: the number of instances belonging to c
3. $|c'|$: the number of instances belonging to c'
4. $|I \cup I'|$: the number of instances of ontologies O and O'.

Figure 2.2(a) shows that, in principle, class extensions are not assumed complete in one data sets. Moreover, two different ontologies usually do not use the same IRIs to denote the same instances, and thus, it is impossible to evaluate $|c \cap c'|$ and $|I \cup I'|$. In this way, cardinalities cannot be exactly determined, and therefore, instance-based matchers have to approximate these cardinalities.

To address this problem we only assume, without loss of generality, that the two sets of instances (I and I') are disjoint but that we have a set of owl:sameAs links $L = \{\langle i, i' \rangle | i \in I \wedge i' \in I'\}$ expressing that the related individuals are actually the same. Although L is correct, it is often incomplete and $|I \cup I'|$ remains over estimated because some same instances may be counted several times.

To that extent, we restrict the computation of cardinalities only on linked instances as follows (see Figure 2.2(b)):

1. $|c \cap c'| = \{(i, i') \in L | (i, c) \in M \wedge (i', c') \in M'\}$
2. $|c| = \{(i, i') \in L | (i, c) \in M\}$
3. $|c'| = \{(i, i') \in L | (i', c') \in M'\}$
4. $|I \cup I'| = |L|$

Furthermore, we also make the unique name assumption on instances that are linked, i.e., which appear in at least one pair of L.

Many measures can be defined for assessing the relation that holds between two classes. For instance, if we consider the subsumption relation, examples of measures are:

1. Inclusion: $m(c, c') = \begin{cases} 1 & \text{if } |c \cap c'| = |c| \\ 0 & \text{otherwise} \end{cases}$
2. Conditional probability: $m(c, c') = \dfrac{|c \cap c'|}{|c|}$
3. Implication intensity: $m(c, c') = \varphi(c \rightarrow c')$ (§2.3)

From the measured values, correspondences can be extracted using simple thresholding, assignment algorithms (greedy, hungarian, etc.) or more elaborate techniques such as [David et al. 2007].

2.3 The implication intensity measure and its entropic version

In this paper, we assess candidate correspondences with a statistical interestingness measure called *implication intensity* [Gras et al. 2008]. It is designed to test quasi-implications $c \rightarrow c'$ and has been already used for ontology matching [David et al. 2007].

Implication intensity, denoted by φ, is the probability that the number of observed counter-examples $|c \cap \overline{c'}|$ to the correspondence $\langle c, c', \sqsubseteq, m\rangle$ is lower that the number X expected under independence hypothesis. The random variable X follows a Poisson law with $\lambda = \frac{|c|.|\overline{c'}|}{|I \cup I'|}$ where $|\overline{c'}| = |I \cup I'| - |c'|$ and $|c \cap \overline{c'}| = |c| - |c \cap c'|$.

The implication intensity is then defined as follows:

$$\varphi(c, c') = 1 - \sum_{k=0}^{|c \cap \overline{c'}|} e^{-\lambda} . \frac{\lambda^k}{k!} \tag{2.1}$$

By using such a measure, we follow the assumption on linked instances: If a pair $(i, c) \notin I$, we assume that i is not an instance of c.

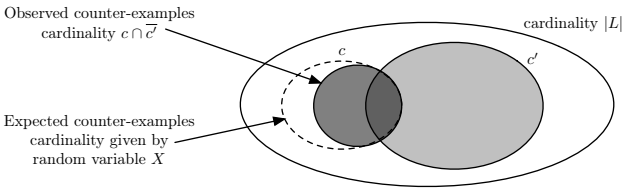


Figure 2.3: Implication intensity among two classes c and c'.

Blanchard et al. [Blanchard et al. 2003] explain that, for large datasets, it is necessary to modulate the value of φ by taking into account the imbalance between $|c \cap c'|$ and $|c \cap \overline{c'}|$

associated with the implication $c \to c'$ and the imbalance between $|c \cap \overline{c'}|$ and $|\overline{c} \cap \overline{c'}|$ associated with the contrapositive $\overline{c'} \to \overline{c}$. They introduce a new measure based on *Shannon's entropy* to non-linearly quantify these differences and to measure the average uncertainty of the random variable. Thus, the weighted version of the implication intensity called the *entropic implication intensity*[1] is given by:

$$\phi(c, c') = (\varphi(c, c').\tau(c, c'))^{1/2} \tag{2.2}$$

[1]The computation of $\tau(c, c')$ is out the scope of this report but it can be found in [Blanchard et al. 2003].

3. Experimental evaluation

We have conducted the experiment in a server of 8 Intel(R) Xeon(R) D-1520, 3.0 GHZ cores, 64GB RAM. For ontology backbone materialization, alignment extraction and to speed up queries, an hybrid database/reasoner called Instance Store (IS) [Horrocks et al. 2004] was built on top of a RDF storage Virtuoso 07.20.3212. We implemented our own ABox reasoner in the IS. It only performs materialization on the basis of both membership (rdf:type) and subsumption (rdfs:subClassOf) triples.

We consider the following datasets for the experiments:

Name	Content	#triples
DBPedia	encyclopedic	3×10^9
Xlore	encyclopedic	24×10^6
GeoLinkedData	geospatial	20×10^9
GeoSpecies	species	2.2×10^6

Table 3.1: Data sets used in these experiments.

DBPedia has been widely used for research and application development because of its breadth and diversity of data: 3 billion RDF triples that are classified with a consistent ontology.

Xlore[1] is a large-scale cross-lingual knowledge base generated from four heterogeneous online wikis: English Wikipedia, Chinese Wikipedia, Hudong Baike and Baidu Baike. It contains 663,740 classes , 56,449 properties and 10,856,042 instances.

GeoLinkedData is a geospatial source that contains a large amount of structured spatial information generated frequently from the Open Street Map (OSM) project[2]. It comprises approximately 20 billion triples[3] and uses a lightweight ontology derived from it.

The GeoSpecies Knowledge Base[4] contains information about 18878 Species, 1650 Families, 217 Orders. Its was designed to help integrate species concepts with species occurrences, gene sequences, images, references and geographical information.

3.1 The extraction process

The alignments are extracted from DBPedia and all other datasets. The process consists of 5 successive steps: LinkSet extraction, LinkSet prunning, Instance type materialization, Consolidation and Results. In the following, we introduce the process through the alignment extraction between Xlore and DBPedia to show how its implementation inevitably requires SPARQL queries customization.

– **LinkSet extraction** The Linkset is a set of correspondences between equivalent instances from two datasets. Since there is not explicit Linkset between Xlore and DBPedia, a SPARQL query (see Query 6.1) has been devised to extract links. In such

[1] `http://xlore.org/`
[2] `http://planet.openstreetmap.org/`
[3] `http://linkedgeodata.org/About`
[4] `https://datahub.io/dataset/geospecies`

query, an Xlore instance and a DBPedia instance are stated as (*sameas*) equivalent if the property value associated with *http://xlore.org/property#hasURL* may be syntactically transformed into a DBPedia resource URI. The Linkset is stored in a two-columns table in the IS.

- **LinkSet prunning** For avoiding additional considerations out of our purpose, the Linkset is pruned to obtain only one-to-one links. Since the Linkset is stored in a two-columns table, we use GROUP BY to group by a chosen column and afterwards the SAMPLE() aggregation function to return randomly one of the values from the other column (see Query 6.2). We applied this strategy symmetrically to both columns. The pruned Linkset is stored in another two-columns table in the IS.

- **Instance type materialization** For each linked instance that make up a link in the pruned Linkset, all its types are materialized. For linked instances from DBPedia, the Query 6.3 is used. For linked instances from Xlore, we use the Query 6.4 which furthermore serves to eliminate the cycles found between classes in the Xlore class hierarchy. The Query 6.4 uses non-standard SPARQL syntax that is provided by the RDF storage. When this subprocess is completed, all this information is stored in a three-column table.

- **Consolidation** Each link of the Linkset is associated with a pair of types (See Query 6.5). Such a pair of types belong to the product of all types of its linked instances. This information is stored a four-column table.

- **Results** From consolidated data, we extract a pairwise classes comparison matrix. Such a matrix contains in each row, quantitative information per pair of classes, one class from DBPedia, the another one from Xlore. For each pair of classes, the quantitative information denotes the size of each class, the size of its intersection and a measure of its relatedness degree. See Queries 6.6, 6.7 and 6.8.

3.2 The linkset

We prune the linkset L to contain one-to-one relationships assuming that they are the most frequent. To validate our assumption we observe the number of one-to-one relationships in the linkset L. To this regard, we adopt the definition given by [Correndo et al. 2012] by which a bundle is the set of instances of a *target* graph reached from a unique instance of a *source* graph. Using a logarithmic scale for the y axis, Figure 3.1 shows the distribution of the frequency of bundles' size. As we can see bundles containing one-to-one relationships are the most frequent.

3.3 The materialized triples

Table 3.2 shows the initial and final size of each graph when S and M are fully classified and materialized respectively. The only exception is Xlore, its materialization (65.7×10^6) was computed only for linked instances since its large classes hierarchy make its full computation not feasible for the hardware used.

The materialization also shows that Xlore data is low quality and inconsistent and therefore, it can bias the alignment as the data does not accurately reflect meaningful both memberships and subsumptions relationships. For example, by Query 6.9 we found non

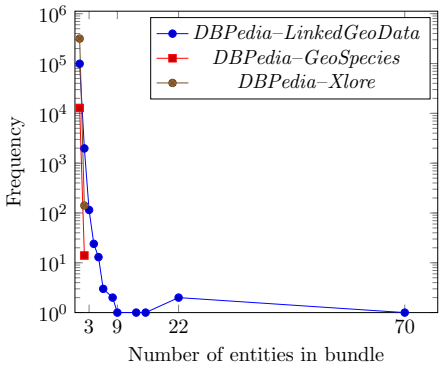


Figure 3.1: Frequency of sameAs bundles by size.

proper types for the instance ⟨http://xlore.org/instance/30085⟩ labeled as "Green sea turtle". Fig 3.2 shows concepts like "Electromagnetism", "Astronomy" or "Corruption" stated as types of such instance.

http://xlore.org/instance/30085	"Green sea turtle"@en	http://xlore.org/concept/356	"Business"@en
http://xlore.org/instance/30085	"Green sea turtle"@en	http://xlore.org/concept/365	"Corporate governance"@en
http://xlore.org/instance/30085	"Green sea turtle"@en	http://xlore.org/concept/7906	"Astronomy"@en
http://xlore.org/instance/30085	"Green sea turtle"@en	http://xlore.org/concept/37747	"Production and manufacturing"@en
http://xlore.org/instance/30085	"Green sea turtle"@en	http://xlore.org/concept/1033	"Corruption"@en
http://xlore.org/instance/30085	"Green sea turtle"@en	http://xlore.org/concept/11	"Organizations"@en
http://xlore.org/instance/30085	"Green sea turtle"@en	http://xlore.org/concept/152	"Countries"@en
http://xlore.org/instance/30085	"Green sea turtle"@en	http://xlore.org/concept/16477	"Industries"@en
http://xlore.org/instance/30085	"Green sea turtle"@en	http://xlore.org/concept/19258	"Pharmaceutical industry"@en
http://xlore.org/instance/30085	"Green sea turtle"@en	http://xlore.org/concept/23	"Electromagnetism"@en
http://xlore.org/instance/30085	"Green sea turtle"@en	http://xlore.org/concept/27386	"Science software"@en
http://xlore.org/instance/30085	"Green sea turtle"@en	http://xlore.org/concept/5222	"Electronic engineering"@en
http://xlore.org/instance/30085	"Green sea turtle"@en	http://xlore.org/concept/7075	"Political organizations"@en

Figure 3.2: Examples of modelling inconsistencies in Xlore

				Materialized	
		S	M	S	M
Datasets	DBPedia	685	11.40×10^6	2376	26.60×10^6
	Xlore	1.60×10^6	4.93×10^6	$23,842 \times 10^6$	65.70×10^6
	GeoSpecies	1989	0.18×10^6	8299	0.94×10^6
	Linkedgeodata	1210	72×10^6	1625	110×10^6

Table 3.2: Materializations.

4. Results

Table 4.1 shows the number of correspondences computed for each alignment. Precision values are only showed for two alignments. Precision for alignment between DBPedia and Xlore is not included because it is too large for manual processing, 62451 correspondences. In all cases, recall values are not computed since reference alignments are not available.

Would you be available

Alignment	correspondences	precision
$DBPedia - LinkedGeodata$	37	0.625
$DBPedia - Geospecies$	20	0.710

Table 4.1: Alignments

Table 4.2 shows the time consumed to extract each alignment. The alignment between DBPedia and Xlore took more time because Xlore has a large classes hierarchy to process in order to materialize type relationships for each linked instance.

Alignment	Time
$DBPedia - LinkedGeodata$	18 min
$DBPedia - Geospecies$	15 min
$DBPedia - Xlore$	11h 40 min

Table 4.2: Time consumed for alignment extraction

5. Conclusions

By materializing typing and subsumption relationships, we are able to extract an instance-based alignment. In addition, we have extended the original use of association rule model provided by [David et al. 2007] to large datasets by using the *entropic implication intensity*. Moreover, our approach has been evaluated obtaining good values of precision for instance-based alignments.

We identify several possibilities for future work. First we plan to extend our approach by considering *rdfs:domain*, *rdfs:range* and *rdfs:subPropertyOf* to materialize additional typing relationships. Moreover, we plan to improve the alignment extraction by pondering another measures. Finally, we plan to implement an algorithm to reduce redundance in the number of extracted correspondences.

6. Appendix

We provide below the various SPARQL queries used in order to extract and manipulate the data in a triple store.

```
SELECT ?dbpediainstance ?xloreinstance
FROM <http://xlore.org> {
?xloreinstance owl:InstanceOf ?xloreclass.
?xloreclass rdfs:label ?label.
?xloreinstance <http://xlore.org/property#hasURL> ?wikiurl.
BIND (iri(replace(str(?wikiurl),
"http://en.wikipedia.org/wiki" ,
"http://dbpedia.org/resource")) AS ?dbpediainstance)
FILTER(REGEX(str(?wikiurl), 'en.wikipedia', 'i')
 && lang(?label) = 'zh')
}
```

Listing 6.1: Linkset extraction

```
SELECT  SAMPLE(instancea) as instance1,
instanceb as instance2,
COUNT(instanceb) as counter2 FROM (
SELECT id_to_iri(instance1) AS instancea,
SAMPLE(id_to_iri(instance2)) AS instanceb,
COUNT(instance1) as counter1
FROM IS.DBA.SAMEAS
WHERE graphname = iri_to_id('<http://xlore.org>')
GROUP BY instance1 ORDER BY counter1 desc)
AS temporal GROUP BY instanceb ORDER BY counter2 DESC
```

Listing 6.2: One-to-one Linkset extraction

```
SELECT DISTINCT ?i ?c  FROM <http://dbpedia.org> {
?i rdf:type/rdfs:subClassOf* ?c.
FILTER (
?i = <%s> &&
!REGEX(str(?c), 'http://dbpedia.org/ontology/Wikidata:') &&
!REGEX(str(?c), 'http://www.ontologydesignpatterns.org/') &&
!REGEX(str(?c), 'http://www.opengis.net/gml/_Feature') &&
!REGEX(str(?c), 'http://www.w3.org/2002/07/owl#Thing') &&
!REGEX(str(?c), 'http://schema.org') &&
!REGEX(str(?c), 'http://wikidata.dbpedia.org') &&
!REGEX(str(?c), 'http://xmlns.com/foaf/0.1/Person')
)
}
```

Listing 6.3: Instance-type materialization for DBPedia

```
SELECT DISTINCT ?instance ?superclass
from <http://xlore.org> {
?class owl:SubClassOf ?superclass
```

```
OPTION (TRANSITIVE, T_DISTINCT,T_NO_CYCLES, T_MIN(0)).
?instance owl:InstanceOf ?class.
FILTER (
?instance = <%s>"
)
}
}
```

Listing 6.4: Instance-type materialization for Xlore

```
SELECT id_to_iri(instance1) as instance1,
id_to_iri(i1.O) as classname1,
id_to_iri(instance2) as instance2,
id_to_iri(i2.O) as classname2
FROM IS.DBA.SAMEAS_ONE_TO_ONE s
INNER JOIN IS.DBA.INSTANCES AS i1
ON s.instance1 = i1.S
INNER JOIN IS.DBA.INSTANCES AS i2
ON s.instance2 = i2.S
WHERE s.graphname = iri_to_id('%s')
AND i1.G = iri_to_id('%s')
AND i2.G = iri_to_id('%s')
```

Listing 6.5: Consolidation

```
SELECT id_to_iri(classname1) AS classname ,
COUNT(DISTINCT instance1)
AS counter FROM IS.DBA.SAMEAS_MATERIALIZATION
WHERE graphname =  iri_to_id(?) GROUP BY classname1
```

Listing 6.6: Computing of DBPedia classes cardinalities

```
SELECT  id_to_iri(classname2) AS classname ,
COUNT(DISTINCT instance2)
AS counter FROM IS.DBA.SAMEAS_MATERIALIZATION
WHERE graphname =  iri_to_id(?) GROUP BY classname2
```

Listing 6.7: Computing of Xlore classes cardinalities

```
SELECT CONCAT(id_to_iri(classname1), '|' ,id_to_iri(classname2))
AS classes,
count(instance1) AS cardinality
FROM IS.DBA.SAMEAS_MATERIALIZATION
WHERE graphname =  iri_to_id(?) GROUP BY classes
```

Listing 6.8: Computing of intersection size between Xlore and DBPedia classes

```
select distinct ?i ?ilabel ?sc ?clabel ?step {
?sc rdfs:label ?clabel.
?c owl:SubClassOf ?sc
```

```
  OPTION (TRANSITIVE, T_DISTINCT, T_NO_CYCLES,
          t_in(?c), t_out(?sc),
          t_step('step_no') as ?step, T_MIN(0)).
?i owl:InstanceOf ?c .
?i ?p ?c.
?i rdfs:label ?ilabel.
filter(?i = <http://xlore.org/instance/30085>
      && lang(?clabel) = 'en'
      && lang(?ilabel) = 'en')
}
order by asc(?step)
```

Listing 6.9: Computing of types for instance "Green sea turtle"

Bibliography

Blanchard, Julien, Pascale Kuntz, Fabrice Guillet, and Régis Gras (2003). "Implication intensity: from the basic statistical definition to the entropic version". In: *Statistical Data Mining and Knowledge Discovery*, pp. 473–485 (cit. on pp. 9, 10).

Correndo, Gianluca, Antonio Penta, Nicholas Gibbins, and Nigel Shadbolt (2012). "Statistical Analysis of the owl: sameAs Network for Aligning Concepts in the Linking Open Data Cloud". In: *Database and Expert Systems Applications - 23rd International Conference, DEXA 2012, Vienna, Austria, September 3-6, 2012. Proceedings, Part II*, pp. 215–230. DOI: 10.1007/978-3-642-32597-7_20 (cit. on pp. 6, 12).

David, Jérôme, Fabrice Guillet, and Henri Briand (2007). "Association Rule Ontology Matching Approach". In: *Int. J. Semantic Web Inf. Syst.* 3.2, pp. 27–49. DOI: 10.4018/jswis.2007040102 (cit. on pp. 9, 16).

Duan, Songyun, Achille Fokoue, Oktie Hassanzadeh, Anastasios Kementsietsidis, Kavitha Srinivas, and Michael J. Ward (2012). "Instance-Based Matching of Large Ontologies Using Locality-Sensitive Hashing". In: *The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I*, pp. 49–64. DOI: 10.1007/978-3-642-35176-1_4 (cit. on p. 6).

Gras, Régis, Einoshin Suzuki, Fabrice Guillet, and Filippo Spagnolo (2008). *Statistical implicative analysis: Theory and applications*. Springer (cit. on p. 9).

Horrocks, Ian, Lei Li, Daniele Turi, and Sean Bechhofer (2004). "The Instance Store: DL Reasoning with Large Numbers of Individuals". In: *Proceedings of the 2004 International Workshop on Description Logics (DL2004), Whistler, British Columbia, Canada, June 6-8, 2004* (cit. on p. 11).

Margara, Alessandro, Jacopo Urbani, Frank van Harmelen, and Henri E. Bal (2014). "Streaming the Web: Reasoning over dynamic data". In: *J. Web Sem.* 25, pp. 24–44. DOI: 10.1016/j.websem.2014.02.001 (cit. on p. 6).

Parundekar, Rahul, Craig A Knoblock, and José Luis Ambite (2010). "Aligning ontologies of geospatial linked data". In: *Workshop On Linked Spatiotemporal Data, in conjunction with the 6th International Conference on Geographic Information Science (GIScience 2010). Zurich, 14th September.(forthcoming 2010)* (cit. on p. 6).

Thor, Andreas, Toralf Kirsten, and Erhard Rahm (2007). "Instance-based matching of hierarchical ontologies". In: *Datenbanksysteme in Business, Technologie und Web (BTW 2007), 12. Fachtagung des GI-Fachbereichs "Datenbanken und Informationssysteme" (DBIS), Proceedings, 7.-9. März 2007, Aachen, Germany*, pp. 436–448 (cit. on p. 6).

Wang, Shenghui, Gwenn Englebienne, and Stefan Schlobach (2008). "Learning Concept Mappings from Instance Similarity". In: *The Semantic Web - ISWC 2008, 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings*, pp. 339–355. DOI: 10.1007/978-3-540-88564-1_22 (cit. on p. 6).

Zong, Nansu, Sejin Nam, Jae-Hong Eom, Jinhyun Ahn, Hyunwhan Joe, and Hong-Gee Kim (2015). "Aligning ontologies with subsumption and equivalence relations in Linked Data". In: *Knowledge-Based Systems* 76.0, pp. 30 –41. ISSN: 0950-7051. DOI: http://dx.doi.org/10.1016/j.knosys.2014.11.022 (cit. on p. 6).